1. What is the use of « label » in SAS et STATA? Why the instruction in the Python does not proceed exactly as in “label” in SAS and STATA? Can you find an equivalent instruction in Python and R?
2. Copy here the image in the slides of the course which signaled some variables had unexpectedly arbitrary value and comment. Copy here the Python code which replaced the observations **-99** by NaN for missing observations for the related ratios.
3. For each explanatory variable in your sample of estimation copy the graphs of distributions and boxplots and report your comments on visual inspection and numbers for skewness, kurtosis, high leverage observations, non-normality for each group, equality of variance, test of equality of means and t-statistics, simple correlation coefficient with the dependent.
4. To which extent normality matters for explanatory variables in this study?
5. Why is it useful to rank the explanatory variables by their correlation coefficient and t-statistics? Why is the order the same when using absolute values of t-statistics or using absolute values of correlation coefficients?
6. Which are the explanatory variables most correlated with the dummy of default? Are there some explanatory variables with a simple correlation with the default dummy below 0.1 in absolute value? What would you expect in this case for predicting default in multivariate analysis?
7. Explain why the t-statistics is the same obtained for 4 different tests with equivalent null hypothesis (cf. slides of the course), including the test of equality of means.
8. Bivariate correlation between explanatory variables: List explanatory variables with simple correlation coefficient with other explanatory variables larger than 0.8? For each case, why is it expected and explained by accounting items included into each ratio?
9. Will you include all of them in your regressions? Explain what are the consequences of highly correlated explanatory variables for regression analysis.
10. Comment the table of the bivariate clouds of points for six variables: yd, total debt and its highly correlated companion ratio, operating income / total assets and its highly correlated companion ratio, and growth of employees. Does the shape of the clouds matches with the simple correlation coefficients?
11. **Comment the items and compare the default output table in the Python code with the alternative single table with a column for each of the estimations: for linear probability model, Logit and Probit when total debt/total assets ratio is the only explanatory variable.**
12. **Explain how you compute the percentage of concordant pairs when the dependent variable is binary.**
13. **Write a single table with a column for each of the estimation: for linear probability model, Logit and Probit with your preferred list of explanatory variables out of the 14 financial ratios. Report at the bottom of each column the value of the area under the ROC curve (AUC) for the sample of estimation, number of observations for each group and after the next question, report below the ROC curve on the sample of validation for each case**

* Benchmarks regressions: tdta, opita, gempl

Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.

Table X reports the estimation results for the three alternative default models using a classic Linear Probability Model, Logit and Probit. Using the Total Debt over Total Asset (tdta), Operating Income to Total Asset (opita) and Gross Employment rate (gempl) as explanatory variable. The estimated coefficients are broadly consistent across models in terms of sign and statistical significance suggesting robustness in the finding, we will go further into the topic in later questions.

Economically the sign of the coefficient also seems to be consistent with theory, tdta being an indicator of leverage a higher value is in fact associated with a higher risk overall, while opita and gempl are indicators of profitability or economic health thus the negative coefficients.

Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.

We also computed the partial effect of the explanatory variables on the probability of default to highlight the economic impact of variable changes on the risk spectrum in a more intuitive probability space than log-odd or z-scores. As Table X reports, for tdta at a probability of default of 25% an increase of 10pp of tdta is associated with an increase of default probability of about 3pp-4pp for the three models. When the probability of default is 50% the same variation of tdta is associated with an increase of the probability of default of 5.14pp (LPM), 6.03pp (Probit) and 6.19pp (Logit). This highlights the nonlinear nature of the Logit/Probit models and is consistent with the theory where some financial changes can have bigger probability default effects for already fragile borrowers.

From now on we will use these models as a benchmark of what estimations and forecasting we can achieve with some economic good sense when it comes to choosing the variable and will compare these results to models using transformed variables (cf: Document A for methodology and step to achieve the transformed variable.)

* Transformed variable regressions: gempl, tdta\_sq, log1p\_opita

Coincidly the transformed variables are the same base variables as what we choose as a benchmark, again to understand the process of selection refer to Document A. This similarity of variables still provides us with the advantages of showing the improvement we can achieve with the same base variable in a different form.

Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, Police, nombre, ligne

Le contenu généré par l’IA peut être incorrect.

It’s not surprising to find the same signs for each variable across the models and the economic meaning stays the same as previously stated. We can still comment on the fact that the significance of the variable associated with total debt over total assets change for LPM, its due to the fact that the transformation to tdta^2 make it harder for a linear model to capture the effect, while being in adequation with the theory relative to the tdta ratio and the convex relationship with the default probability from previous credit risk studies.

Concerning the marginal effect of the variables on the default probability, the same kind of relationship can be made. The leverage is still the main indicator of default risk and while we could integrate ourselves about the relevance of using other variables in this case, we can make the assumption that considering positive economic health factors in addition increase the accuracy of the model by capturing at least a part of the complex relationship of financial ratio. It’s plausible that some high leverage firms never default because of an otherwise healthy situation.

1. **For the sample of VALIDATION (precise its number of observations: n0= for non-defaulting firms, n1=… for defaulting firms), compute the AUC, area under the ROC curve, for your** preferred list of explanatory variables for linear probability model, Logit and Probit. (You are free to add any other statistics on the sample of validation which you think they are interesting). Compare the AUC of the sample of validation to the AUC of the sample of estimation for each model.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

The Table X and Y show the models performances metrics such at the AUC, the Accuracy (define as the true diagnosis over total diagnosis) and the confusion matrix. Let’s add to the mix the ROC curves to visualize the two table and compare with the ROC curve of the estimation samples.

Une image contenant texte, diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.Une image contenant texte, diagramme, ligne, Tracé

Le contenu généré par l’IA peut être incorrect.

As Table X and Y suggest the explanatory variables give similar results, when comparing for the same sample the ROC curve yields similar shape and characteristics. Regarding each specification, the AUC (therefore ROC curve) show better metrics for the training sample is to be expected, most of the time the models will perform better on the data they were trained on (its not necessarily the case, some model might be better at generalizing but the noise on small sample like ours make the AUC scores too dependent on initial sampling). It seems that the variable we have chosen at first with an economic and a simple correlation analysis (benchmark) performs very well. Again, the similarity in the benchmark and the transformed variables we choose are mainly due to chance, in fact after some test it seems that it was one of the best “naïve” choices of variable, it just happens that the systematic method we used for the transformed variable have yield to the same base variable combination.

Too determine further if the two variables selection are different despite similar performances lets run the standardized Pearsons residuals.

1. **Comment and compare the plots of the distributions of standardized Pearson residuals of the linear probability model, for the Logit model and for the Probit model. Write their formula and then explain for which reason they do not fit a normal law?**

Une image contenant texte, capture d’écran, Tracé, ligne

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, ligne, capture d’écran, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant capture d’écran, ligne, Tracé, texte

Le contenu généré par l’IA peut être incorrect.

The graph X,Y and Z report the standardized Pearson residuals for the three models of the transformed variables. The standardized Pearson residuals can be obtained by doing:

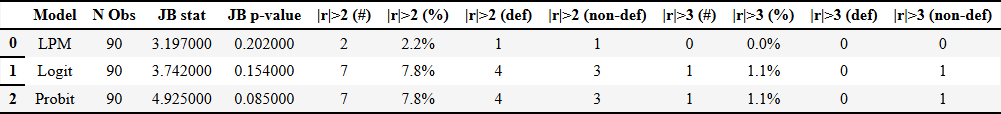
Where is the dummy variable for default or non-default and the residuals of forecasted probability.

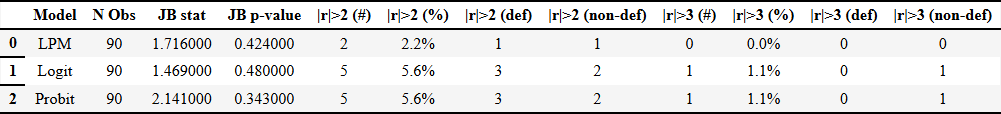
This diagnostic tool allows us to visualize the repartition of the residuals of forecasted probability for probability models (like our LPM, Probit, Logit). Usually, we can expect for a well specified model that the residual will follow a normal law, with the residual center at 0 and no value past the -1.96:1.96 threshold.

As the graph X, Y and Z report, the standardized Pearson residuals do not follow perfectly a normal law. We can associate the non-normality in our case with a difficulty for the model to distinguish either what we can call type 1 outliers (right side) with good indicators but an observed default (surprise defaulting) and type 2 outliers (left side) with bad indicators but an observed non-default (resilient firm). It does not necessarily say that the model is bad, we will talk about the nature of the outlier in the next question, but it is rather kind of expected in a realistic scenario where even if a firm have a lot of bad indicators the complex economic relationship leading to a default or not can produce such outliers.

1. **Do you find outliers for standardized Pearson residuals? Which side of extreme values are providing these outliers for defaulting firms and for non-defaulting firms?**

Now lets take a look at the standardized Pearson residuals for both our benchmark models and transformed variables models.





As the table X and Y report the p-value for the Jarque-Bera test for the residuals are all above 0.05 so we cannot reject normality strongly for any models. For both specification the only big outliers (|r|>3) and most of the other outliers (|r|>2) are type 1 outliers (surprising default yd = 1) meaning all models tends to underestimate more the probability of default for non-default looking firm (false negative) than overestimating non-default probability for default looking firm (false positive). This observation matches table Z from Q14.

We can add additionally that we have less outliers for the probit and logit models in the case of transformed variables, we can assume the transformed variables are marginally better at capturing limit cases at least for this sample and splitting.

1. **For a loss function of the type 1 and type 2 errors, what would be relative weight for a private banker ? How one can take into account this loss function for selecting between scoring models taking into account the ROC curve of each model, in particular if the ROC curve intersect (locally one is over the other and conversely).**

A private banker would follow the loss function given by:

While choosing the threshold that minimize Loss we can note that:

Given a Loss value we can compute (for example the minimal Loss):

The equation of the loss function in the ROC space. We can now graphicly estimate the treshold s\* as the tangent point between the ROC curve and the lowest possible iso-loss. This is equivalent to choose a treshold s\* such that the marginal cost of Type I error is equal to the marginal cost of Type II error.

1. **Now, you got a job for providing credit to private firms. What weight would you give to the score of this firm for giving or not credit to this firm with respect to its business and financial plans and financial analysis?**

If we would have to choose a threshold given our models, we first need to fixe an LGD and a margin . We decided to put LGD = 0.6 and margin = 0.1 to compute the optimal threshold of our models.

First let’s note the theorical optimal threshold in closed conditions :

For the probability of default for a firm:

We lend if:

In our case with LGD = 0.6 and margin = 0.1, = 0.1429, wich mean if a firm have less than 14.29% of default probability we lend and we do not lend if the probability is greater than 14.29%.

Now let’s compute the threshold in experimental conditions with our models:

Une image contenant texte, ligne, diagramme, Tracé

Le contenu généré par l’IA peut être incorrect.

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

As reported by Table X the threshold for Logit is 12.20% meaning we won’t lend to firm with a greater probability of default than 12.20% according to the logit model with the transformed variables. This number is less than the closed threshold but keep in mind that the closed threshold is 100% cost structure dependant whereas the experimental threshold is partially driven by the data (still a part that is cost structure related). We can deduce that compare to the closed threshold that assume perfect calibration a lower threshold for experimental models is since the model can misclassified borrower around the threshold making lending money at let’s say 14.29% more risky than the theory. To minimize type I error (most costly) the model chooses to increase Type II error (less costly). This behaviour is expected for two reasons mainly, first the Logit and Probit models produce more extreme probability of default due to the distribution of the models, making mid-range firm more prompt to Type I errors, secondly the cost structure chosen make a type I error six time as costly as a Type II error, if we choose to shift the margin closer to the LGD (or vice versa) we could obtain different kinds of closed versus empirical threshold.

At the end the key value here for the performance evaluation of the models are the Min expected Loss, for our case the minimal expected loss ranges from [3.8 to 3.6] for 90 observations, equivalent to 4.22% and 4%. We can expect for the logit model with the transformed variable [gempl, tdta\_sq, log1p\_opita] to be able to lend with a default rate of 4%. A quick computation for 1$ lends to 100 firms gives us 4 defaulting firms giving us back 1.6$ and 96 firm giving us back 105.6$ for a net profit of 7.2%.

1. **Question on the dummy trap: create or generate a dummy equal to one for non-default firms named ynd, using the command “ynd=1-yd”. Then do the following OLS regressions with TDTA (total debt over total assets) as the dependent variable and comment all these intriguing results:**
2. **Tdta regressed on yd, ynd with common intercept**
3. **Tdta regressed on yd and common intercept (on the slides)**
4. **Tdta regressed on ynd and common intercept**
5. **Tdta regressed on yd, ynd with common intercept and the restriction that the sum of parameters of yd and ynd are equal to zero.**

Une image contenant texte, capture d’écran, Police, nombre

Le contenu généré par l’IA peut être incorrect.

1. Explain practical changes you did with respect to the given STATA, SAS, PYTHON or R code (if any).